Institute for Data Driven Dynamical Design ID_4

NSF-OAC# 2118201

AI-Ready Data: Knowledge Extraction from Laboratory Notebooks

Elizabeth Jones[1], Joel Pepper[2], Kyle Langlois[3], Jacob Furst[3], Fernando Uribe-Romo[3],
Jane Greenberg[2], David E. Breen[2]
[1]Northeastern University, [2]Drexel University, [3]University of Central Florida

Drexel University Metadata Research Center
College of Computing & Informatics

UCF Department of Chemistry
University of Central Florida

Motivation

- Within chemistry and material science, experimental protocols are often recorded in handwritten notebooks. Although these notebooks are rarely studied directly, they contain valuable information on what differentiates failed and successful experiments.
- A significant barrier to performing computational analysis of lab notebooks is digitizing their contents into a structured, machine-readable form.
- In this study, we are exploring information extraction methods to make analog lab notebook data computationally ready and determine if there exist any patterns in the data that separate successful and unsuccessful experiments.
- We look specifically at notebooks detailing synthesis experiments for metal-organic frameworks (MOFs), compounds that are of interest for their possible applications to gas purification, gas separation, water remediation, and as conducting solids and supercapacitors.

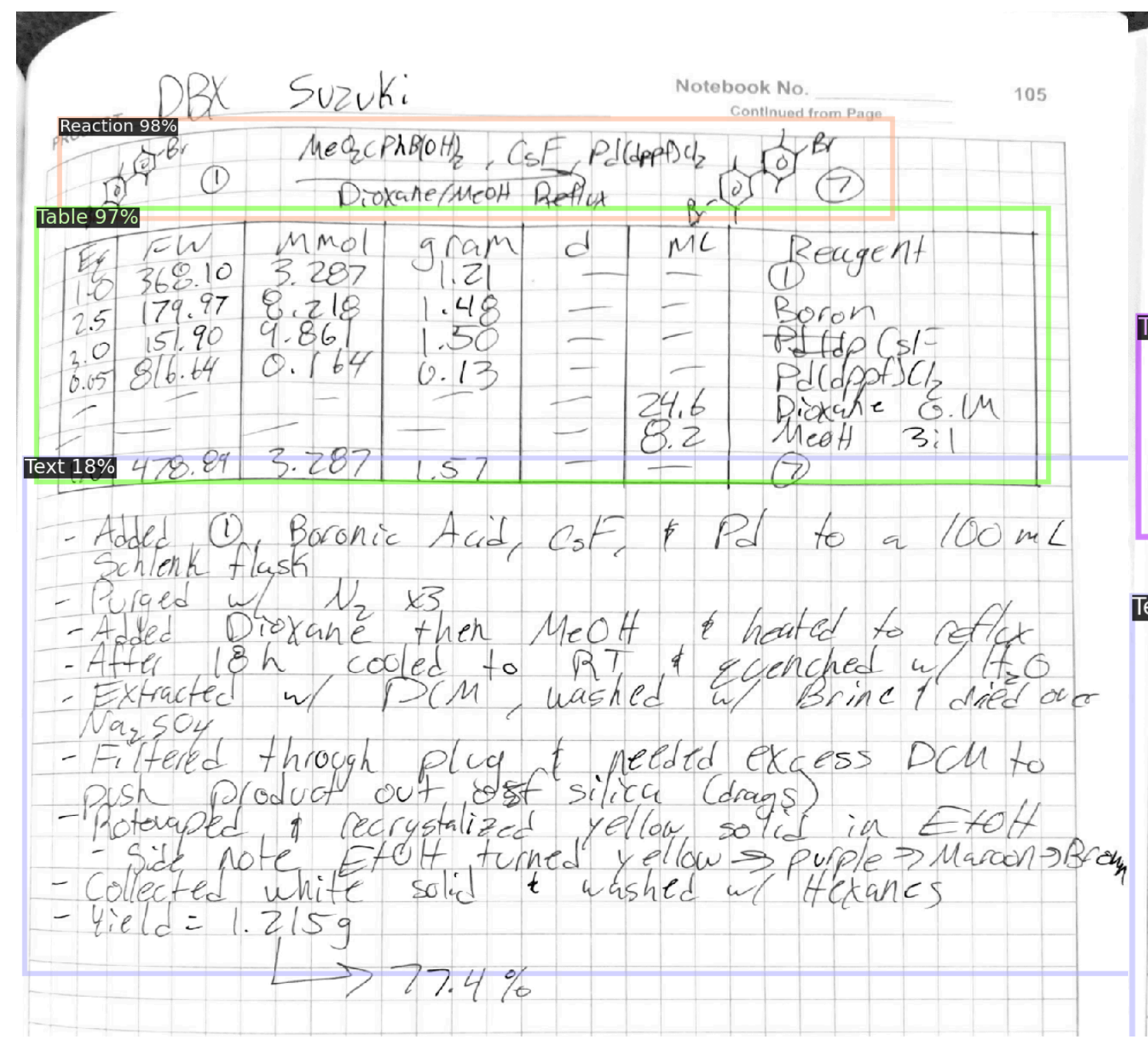
Goals

Overall Goal: digitize and structure handwritten notebook data for AI analysis.

- Investigate approaches for document segmentation, optical character recognition, and text tokenization for methods to extract a variety of information from the scanned lab notebooks.
- Extract data to make accessible or usable to MOF scientists so they can benefit from previous research, and use data to train AI models for scientific inquiries.

Methods and Procedures

- Conduct a literature review to find pre-existing computational tools that could be used or modified for segmentation of the pages, optical character recognition, or text tokenization.
- Segment notebook pages into three main sections: Table, Text, and Reaction. Using the "Makes Sense" annotation tool (makesense.ai), I drew bounding boxes and labeled each element. Then saved all the labeled training data to a CSV file.
- For this project, we chose to use the Facebook AI Research's Detectron2 object detection library due to its many flexible and robust capabilities.

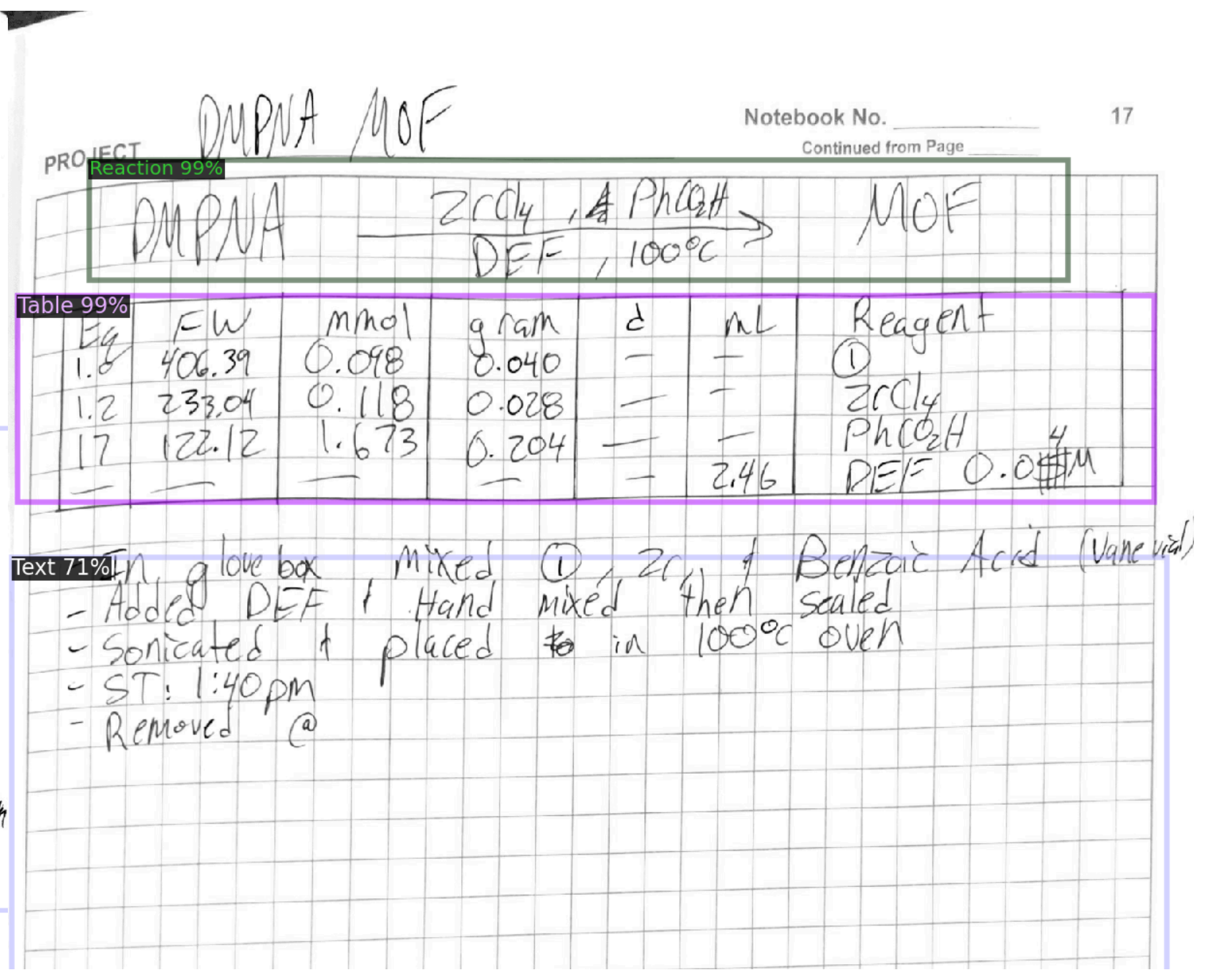


DBX Suzuki
Notebook No.
Continued from Page
105

① MeO2C PhB(OH)2, CsF, Pd(dppf)Cl2 / Dioxane/MeOH Reflux ⑦

Eq	FW	mmol	gram	d	mL	Reagent
1.0	368.10	3.287	1.21	—	—	①
2.5	179.97	8.218	1.48	—	—	Boron
3.0	151.90	9.861	1.50	—	—	~~Pd(dp~~ CsF
0.05	816.64	0.164	0.13	—	—	Pd(dppf)Cl2
—	—	—	—	—	24.6	Dioxane 0.1M
				—	8.2	MeOH 3:1
	478.84	3.287	1.57	—	—	⑦

- Added ①, Boronic Acid, CsF, & Pd to a 100 mL Schlenk flask
- Purged w/ N_2 x3
- Added Dioxane then MeOH & heated to reflux
- After 18h cooled to RT & quenched w/ H_2O
- Extracted w/ DCM, washed w/ Brine & dried over Na_2SO_4
- Filtered through plug & needed excess DCM to push product out ~~of~~ silica (drags)
- Rotevaped & recrystalized yellow solid in EtOH
  - Side note EtOH turned yellow → purple → Maroon → Brown
- Collected white solid & washed w/ Hexanes
- Yield = 1.215g → 77.4%



PROJECT DMPNA MOF
Notebook No.
Continued from Page
17

DMPNA $\xrightarrow[\text{DEF, 100°C}]{ZrCl_4,\ PhCO_2H}$ MOF

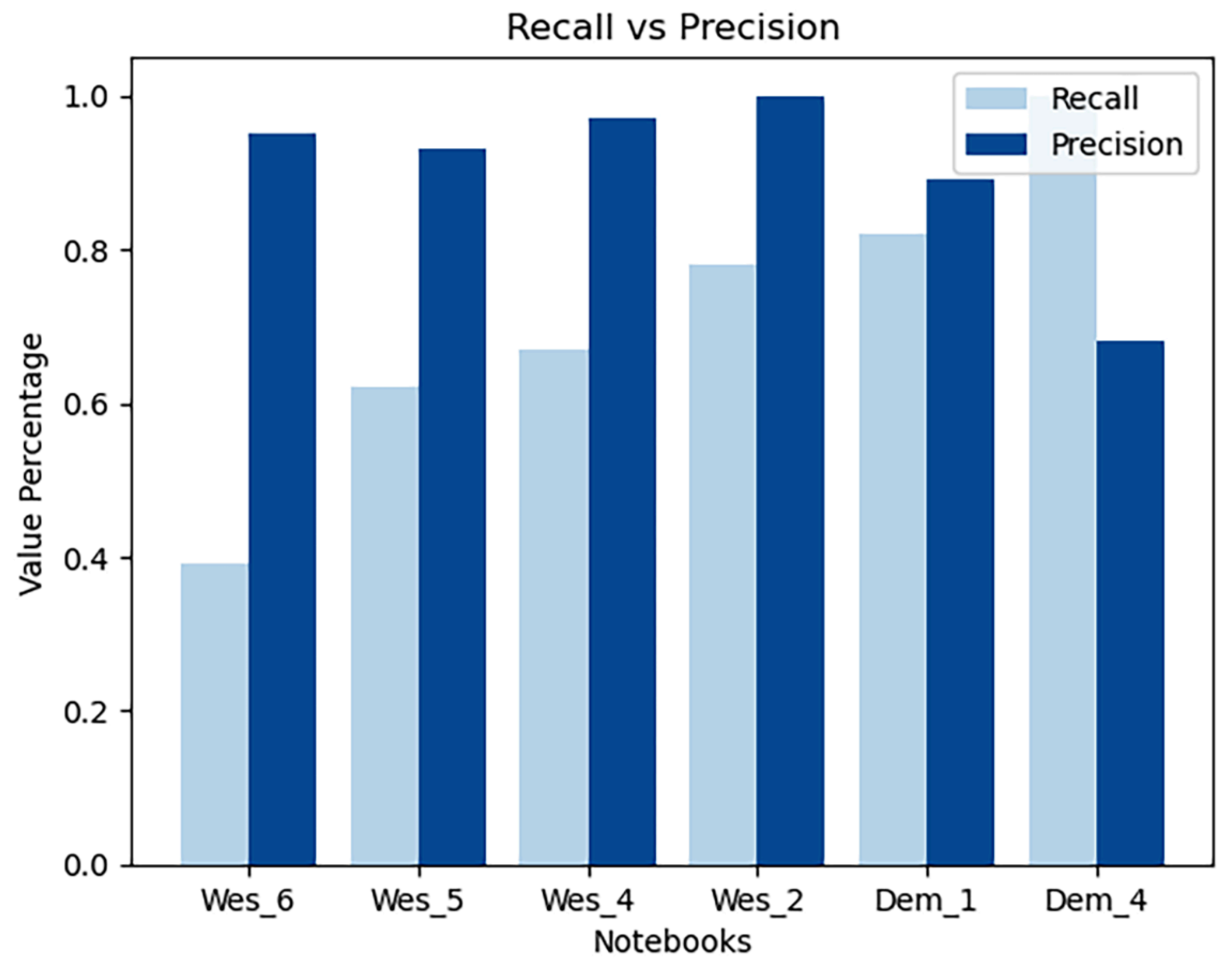
Eq	FW	mmol	gram	d	mL	Reagent
1.0	406.39	0.098	0.040	—	—	①
1.2	233.04	0.118	0.028	—	—	ZrCl4
17	122.12	1.673	0.204	—	—	PhCO2H
				—	2.46	DEF 0.04M

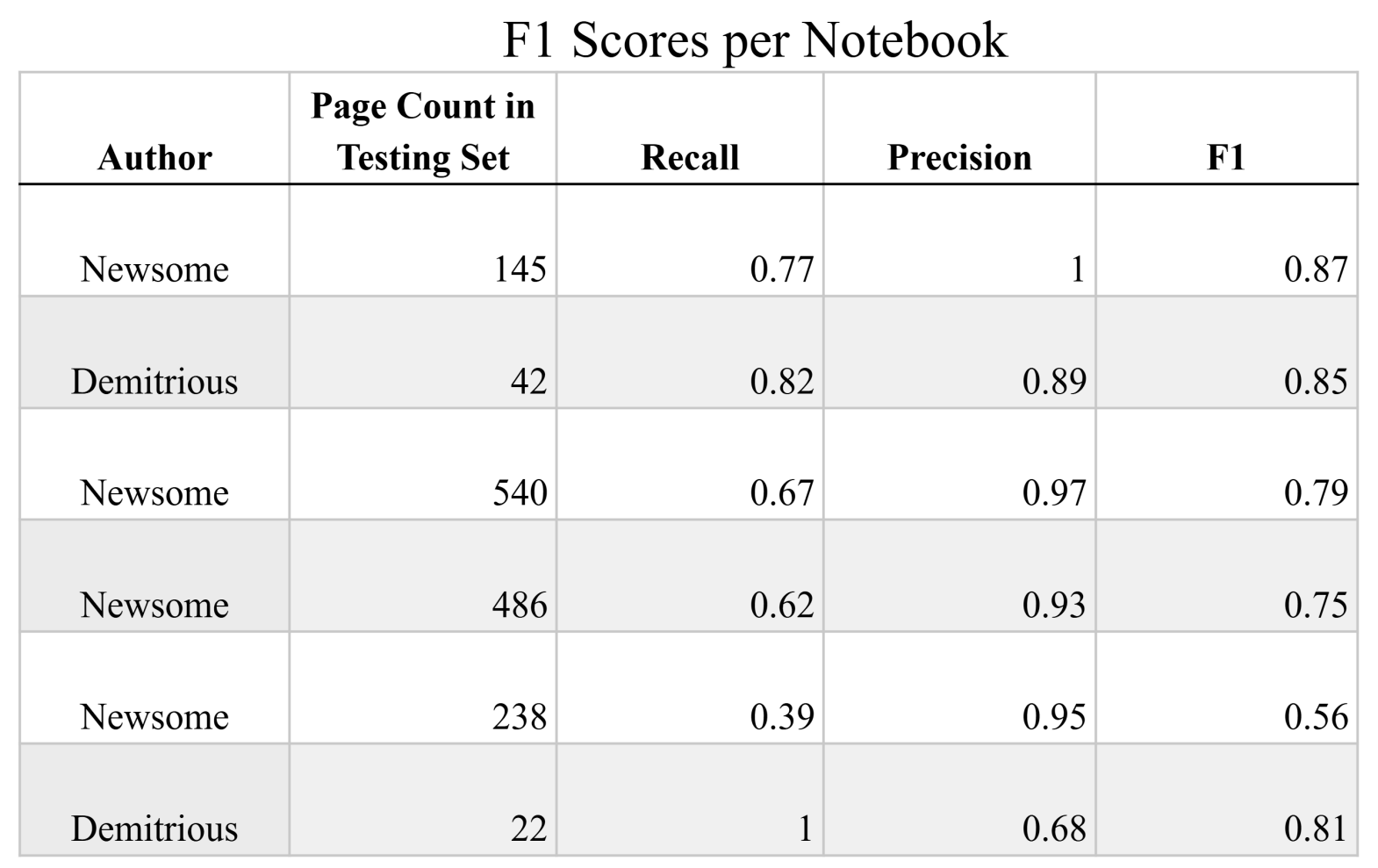
In glove box mixed ①, Zr, & Benzoic Acid (vane vial)
- Added DEF & Hand mixed then sealed
- Sonicated & placed in 100°C oven
- ST: 1:40pm
- Removed @

- Train the object detection software on hand-labeled notebook pages, and run tests on un-labeled pages to assess model accuracy.
- Analyze the results after testing the remaining pages from the notebooks, make necessary adjustments, increase the training set, and continue to train our object detection model on more notebooks.

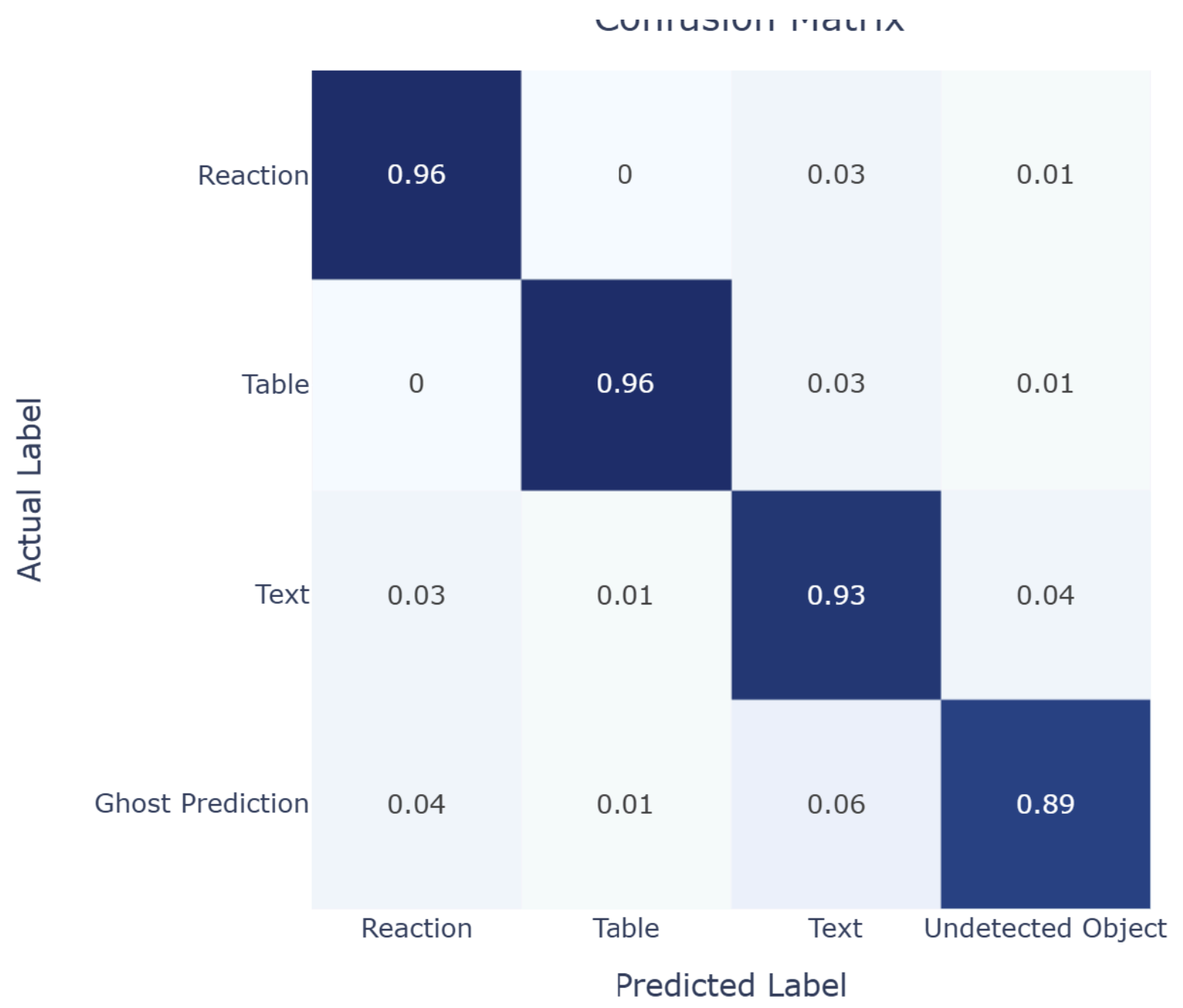
Results

- After training our model on two notebooks from two different authors, we tested on the remaining pages of those notebooks as well as four more notebooks by the same authors.
- I took note of results from each page and calculated the True Positive, False Positive, False Negative, Recall, Precision, and F1 values per notebook and object type.


Recall vs Precision


F1 Scores per Notebook

Author	Page Count in Testing Set	Recall	Precision	F1
Newsome	145	0.77	1	0.87
Demitrious	42	0.82	0.89	0.85
Newsome	540	0.67	0.97	0.79
Newsome	486	0.62	0.93	0.75
Newsome	238	0.39	0.95	0.56
Demitrious	22	1	0.68	0.81



Future Work

- Improve model accuracy by adding more notebooks to the training data.
- Test the model using different cut-offs for object detection confidence and determine an optimal cut-off percentage.
- Perform more granular data extraction using techniques such as OCR and LLMs, and develop a structured representation of the data that can be fed to ML tools.
- Apply unsupervised clustering on structured representations of the notebooks, specifically to look for patterns in the data more thoroughly to determine differences between failed and successful experiments.

References

[1] Chen et al., 2011, "Table Detection in Noisy Off-Line Handwritten Documents"
[2] Dozias et al., 2018, "Smart Pens to Assist Fibre Optic Sensors Research"
[3] Franco-Gaona et al., 2024, "Towards the Automatic Extraction and Annotation of Information Elements from Handwriting Notes"
[4] Gaona et al., 2020, "Extracting Information Objects from Handwriting Laboratory Notes"
[5] Weir et al., 2021, "Chempix: Automated Recognition of Hand-Drawn Hydrocarbon Structures Using Deep Learning"

We acknowledge the original authors of the lab notebooks: Wesley Newsome, and Demitrious Vazquez-Molina